WHAT IS CLAIMED:

1. A method for optimizing a database of sample records for the training and testing of a prediction algorithm, comprising:

defining a set of one or more distirbutions of the database records onto respective training and testing subsets;

using the defined set of distributions to train and test a first generation set of one or more prediction algorithms and assigning a fitness score to each;

feeding the set of prediciton algorithms to an evolutionary algorithm which generates a set of one or more second generation prediction algorithms and assigns a fitness score to each; and

continuing to feed each generational set of prediction algorithms to the evolutionary algorithm until a termination event occurs;

where said termination event is at least one of a prediciton algorithm is generated with a fitness score equalling or exceeding a defined minimum value, the maximum fitness score of sucessive generational sets of prediction algorithms converging to a given value, and a certain number of generations having been generated.

2. The method of claim 1 characterised in that it comprises the following steps

generating a population of prediction algorithm each one of them is trained and tested according to a different distribution of the records of the data set in the complete database onto a training data set and a testing data set;

each different distribution being created by a random or pseudo-random distribution;

each prediction algorithm of the said population is trained according to its own distribution of records of the training set and is validated in a blind way according its own distribution on the testing set;

a score reached by each prediction algorithm is calculated in the testing phase representing its fitness;

an evolutionary algorithm being further provided which combines the different models of distribution of the records of the complete data set in a training and in a testing set which sets are represented each one by a corresponding prediction algorithm trained and tested on the basis of the said training and testing data set according to the fitness score calculated in the previous step for the corresponding prediction algorithm;

the fitness score of each prediction algorithm corresponding to one of the different distributions of the complete data set on the training and the testing data sets being the probability of evolution of each prediction algorithm or of each said distribution of the complete data set on the training and testing data sets.

Repeating the evolution of the prediction algorithm generation for a finite number of generations or till the output of the genetic algorithm converges to a best solution and/or till the fitness value of at least some prediction algorithm related to an associated data records distribution has reached a desired value;

Setting the data records distribution for the best solution as the optimized training and testing subsets for training and testing prediction algorithm.

3. A Method according to one or more of the preceding claims characterised in that to each record of the data set a distribution variable is associated which is binary and has at least two status, one of this two status being associated with the inclusion of the record in the training set and the other in the testing set.

4. A Method according to one or more of the preceding claims characterised that the prediction algorithm is an artificial neural network.

5. A Method according to one or more of the preceding claims 1 to 4, characterised in that the prediction algorithm is a classification algorithm.

6. A Method according to one or more of the preceding claims characterised in that once an optimum distribution has been computed, the optimised training data subset is made equal to a complete data set being the individuals included in the training subset distributed onto a new training set and onto a new testing set each one having about the half of the records of the original optimized training set, while the originally optimized testing set is used as a third data subset for validation purposes.

7. A Method according to one or more of the preceding claims, characterised in that the distribution of the data of the originally optimized training set onto the new training and new testing set is optimized by means of a pre-processing phase according to claims 1 to 4.

8. A Method according to claim 1, in which the different choices of the structure of the training and of the testing data subsets consist in different selections of the number of input variables of the data records of the database, which selections consist in leaving out at least one, preferably two or more variables from the entire input variable set forming each record, the records of the data base comprising a certain number of known input variables and a certain number of known output variables.

9. A Method according to claim 8, characterised by the following steps:

Defining a distribution of data from the complete data set onto a training and onto a testing data set;

Generating a population of different prediction algorithm each one having a training and/or testing data set in which only some variables have been considered among all the original variables provided in the data sets, each one of the prediction algorithm being generated by means of a different selection of variables.

Carrying out learning and testing of each prediction algorithm of the population and evaluating the fitness score of each prediction algorithm.

Applying an evolutionary algorithm to the population of prediction algorithms for achieving new generations of prediction algorithm;

For each generation of new prediction algorithms representing each one a new different selection of input variable, the best prediction algorithm according to the best hypothesis of input variables selection is tested or validated.;

A fitness score is evaluated and the prediction algorithms representing the selections of input variables which have the best testing performances and the minimum input variables are promoted for the processing of the new generations.

10. A Method according to claim 8 or 9, characterised in that the pre-processing phase for selecting the most predictive input variables is carried out in combination to a Method according to one or more of the claims 2 to 9.

11. A Method according to claim 10, characterised in that the database subjected to a pre-processing phase of input variable selection according to claims 8 to 9 is a training subset and a testing subset processed with the method according to claim 2 to 8 for data records distribution optimization.

12. A Method according to one or more of the preceding claims 1 to 11, characterised in that the complete database the distribution of the records of which has to be optimized with a Method according to claims 2 to 8 has data records having a selected number of input variables, the selection being carried out with a Method according to claims 8 and 10.

13. A Method according to one or more of the preceding claims characterised in that the pre-processing phases for optimizing the distribution of the records on a training and a testing subset and of Selecting the most predictive input variables, may be carried out alternatively one to the other several times.

14. A method according to one or more of the preceding claims characterised in that the evolutionary algorithm is a genetic algorithm with the following evolutionary rules:
An average health value of the population is computed as a function of the fitness values of each single individual in the population.

Coupling, kind of recombination of genes and mutation of genes is carried out in a differentiated manner depending o the comparison between the fitness of each individual of the couple and the average health value of the entire population to which the individuals belong.

Individuals having a fitness value lower or equal to the average health of the entire population are not excluded from the creation of new generations but are marked out and enters a vulnerability list;

The number of subject entered in the vulnerability list defining the number of possible marriages.

15. A Method according to claim 14 in which for coupling purposes and for generation of children both parent individuals must have a fitness value close to the average health of the entire population.

16. A Method according to claims 14 and 15, characterised in that each couple of individuals can generate individuals having a fitness different from the average health, so called offsprings if the fitness of one them, at least is greater than the average fitness, the offsprings of each marriage occupying the places of subjects entered in the vulnerability list and are marked out, so that a weak individual can continue to exist through his own children

17. A Method according to claims 14 to 16, characterised in that coupling between individuals having a very low fitness value and a very high fitness value are not allowed.

18. A Method according to claims 14 to 17, characterised in that the following recombination rules of the genes of the parents individuals coupled are considered in the case the parents individuals have not common genes:

the health of father and mother individuals are greater than the average health of the entire population; the crossover is a classical crossover according to which the genes of the father and of the mother individuals are substituted one with the other starting from a certain crossover point.

the health of father and mother individuals are lower than the average health of the entire population; In this case the two children are formed through rejection of the parents genes they will receive by the crossover process;

the health of one of the parents is less than the average health of the entire population while the health of the other parent is greater than the average health of the entire population; In this case only the parents whose health is greater than the average health of the entire population will transmit their genes, while the genes of the parent having an health lower than the average health of the entire population are rejected.

19. A Method according to claim 18, characterised in that genes rejection consist in modifying the status of the genes variable of the individuals from one to a following status level defined for this genes (variable).

20. A Method according to claims 18 or 19, characterised in that a modified crossover of the genes of the parents individuals is carried out when the parents individuals has part of the genes that coincide, this modified crossover provides for

generating and offspring in which the genes selected for crossover are the most effective ones of the parents.

21. A Method according to one or more of the preceding claims 14 to 20 in which the individuals are the different prediction algorithm representing a corresponding different initial random distribution of data records onto the testing and the training data set and the genes consist in the binary status variable of association of each record to the training and to the testing subset.

22. A Method according to one or more of the preceding claims 14 to 20 in which the individuals are the prediction algorithms each one representing a different training and testing data set, the difference residing in a different selection of input variables for each different training and testing subset, and the genes consist in the different selection variable which is provided for each input variable in the different training and testing subsets, the above mentioned selection variable being a parameter indicating the presence/absence of each corresponding input variable in the records of each data set.

23. A method according to one or more of the preceding claims characterized in that it is in the form of a software program comprising instructions executable by a CPU, the software program being stored in a memory to which the CPU can access.

24. A software program stored on a memory device, the said software program consisting in the method according to one or more of the preceding claims in the form of a executable instructions of a CPU or of a computer system.

25. A system for carrying out a method according to one or more of the preceding claims comprising an apparatus or device for generating an action of response which is autonomously, i.e. by itself, chosen among a certain number of different kinds of actions of response stored in a memory of the apparatus or autonomously generated by the apparatus basing the said choice of the kind of action of response on the interpretation of data collected autonomously by means of one or more sensors responsive to physical entities or which are fed to the apparatus by means of input means, the said interpretation being made by means of a prediction algorithm in the form of a software saved in a memory of the said apparatus and being carried out by a central processing unit, characterized in that the apparatus being further provided with

means for carrying out a training and testing phase of the prediction algorithm by inputting to the said prediction algorithm data of a known database in which input variables of the input data representing the physical entities able to being sensed by the apparatus through the one or more sensors and/or able to be fed to the apparatus by means of the input means are univoquely correlated to at least one definite kind of action of response among the different kinds of possible action of response, the said means for carrying out the training an testing being in the form of a training and testing software saved in a memory of the apparatus, the said training and testing being carried out by means of a method according to one or more of the preceding claims 1 to 22, the said training and testing software program being the said method of training and testing in the form of a software program or instructions.

26. The system according to claim 25, characterized in that it is a system for sound or vocal recognition comprising input means responsive to acoustic waves, a processing unit connected to the input means responsive to acoustic waves, at least a memory in which a software program is stored the said program being in the form according to claims 23 or 24 and comprising coded instructions for enabling the processing unit to carry out a method according to one or more of the preceding claims 1 to 22, a further or the same above mentioned memory in which a dataset of known data records is stored or can be stored and/or input means for storing in the further or the said above mentioned memory a dataset of known data records.

27. The system according to claim 25 or 27, characterized in that it is a system for image recognition, the input means being responsible to electromagnetic waves, the system being able to recognize the shape of an object generating or reflecting electromagnetic waves, and/or the distance and/or the identity of the object.

28. The system according to claims 26 or 27, characterized in that the database of known data records comprises acoustic signals emitted by one or more objects or one or more living beings making part of the typical environment in which the device has to operate or the data relating to one or more images of one or more objects or one or more living beings making part of the typical environment in which the device has to operate to which are univoquely correlated to corresponding known kind, and/or identity and/or meaning of objects to which the said acoustic signals or image data are related and/or from which the said acoustic signals or image data are generated.

29. The system according to one or more of the preceding claims 27 or 28, characterized in that it is a specialized system for image pattern recognition having artificial intelligence utilities for analyzing a digitalized image, i.e. an image in the form of a array of image data records, each image data record being related to a zone or point or unitary area or volume of a two or three dimensional visual image, so called pixel or voxel of a visual image, the said visual image being formed by an array of the said pixels or voxels and utilities for indicating for each image data record a certain quality among a plurality of known qualities of the image data records;

the system having a processing unit as for example a conventional computer, a memory in which an image pattern recognition algorithm is stored in the form of a software program which can be executed by the processing unit,

a memory in which a certain number of predetermined different qualities which the image data records can assume has been stored and which qualities has to be univoquely associated to each of the image data records of an image data array fed to the system,

input means for receiving arrays of digital image data records or input means for generating arrays of digital image data records from an existing image and a memory for storing the said digital image data array,

output means for indicating for each image data record of the image data array a certain quality chosen by the processing unit in carrying out the image pattern recognition algorithm in the form of the said software program;

the image pattern recognition algorithm is a prediction algorithm in the form of a software program, which prediction algorithm is further associated to a system being further provided with a training and testing software program,

the system is able to carry out training and testing according to the method of of one or more of the preceding claims 1 to 22,

the method is provided in the system in the form of the training and testing software program,

a database being also provided in which data records are contained univoquely associating known image data records of known image data arrays with the corresponding known quality from a certain number of predetermined different qualities which the image data records can assume.

30. A method for producing a microarray for genotyping operations, the said method comprising the steps of defining a certain number of theoretically relevant genes or alleles or polymorphisms considered relevant for a certain biologic condition like a tissue structure, a pathology or the potentiality of developing a pathology or an anatomic or morphologic feature;

a) providing a database of experimentally determined data in which each record relates to a known clinical or experimental case of a sample population of cases and which records comprise a certain number of input variables corresponding to the presence/absence of a certain predetermined number of polymorphisms and/or mutations and/or equivalent genes of a certain number of theoretically probable relevant genes and one or more related output variables corresponding to the certain biological or pathologic condition of the said clinical and experimental cases of the sample population;

characterized by the following further steps

b) determining a selection of a reduced number of the certain predetermined number of polymorphisms and/or genes by testing the association of the said genes or polymorphisms and the biological or pathological condition by means of mathematical tools applied to the database.

c) The said mathematical tools comprise a so called prediction algorithm such as a so called neural network;

and the further steps are carried out of

d) dividing the database in a training and a testing dataset for training and testing the prediction algorithm;

e) defining two or more different training dataset each one having records with a reduced number of the input variables which reduced number of input variables is obtained by excluding one or more input variables from the originally defined number of input variables, while for each record the reduced number of input variables of the corresponding training set has at least one input variable which is different from the input variables of the reduced number thereof of the other training datasets, each different input variable consisting in a different gene or a different polymorphisms and/or a different mutation and/or a different functionally equivalent gene thereof of the originally considered genes or polymorphisms and/or mutations and/or functionally

equivalent genes thereof considered theoretically potentially relevant for the biologic or pathologic condition;

f) training the prediction algorithm with each of the different training sets defined under point e) for generating a first population of different prediction algorithm which are divided into two groups of mother and father prediction algorithms and testing the said prediction algorithms with the associated testing set;

g) calculating a fitness score or prediction accuracy of each father and mother prediction algorithms of the said first population by means of the testing results;

i) providing a so called evolutionary algorithm such a genetic algorithm and applying the evolutionary algorithm to the first population of mother and father prediction algorithms for achieving new generation of prediction algorithms whose training and testing dataset comprises records whose input variables selections are a combination of the input variable selections of the records of the training and of the testing datasets of the first or previous population of father and mother prediction algorithms according to the rules of the evolutionary algorithm;

j) for each generation of new prediction algorithms representing each new variant selection of input variables, the best prediction algorithm according to the best hypothesis of input variable selection is tested or validated by means of the testing dataset;

k) a fitness score is evaluated and the prediction algorithms representing the selections of input variables which have the best testing performance with the minimum number of input variables utilized are promoted for the processing of new generations;

l) repeating the steps i) to k) until a predetermined fitness score defined as best fit of the prediction algorithm and a minimum number of input variables has been reached;

m) defining as the selected relevant input variables i.e. as the relevant genes or polymorphisms and/or of mutations and/or of functionally equivalent genes thereof the ones related to the input variables of the selection represented by the prediction algorithm having both at least the predetermined fitness score and also the minimum number of selected input variables.

31. A method according to claim 30, characterized in that an optimization of the distribution of the records of the original database in a training and in a testing database

is carried out as a pre processing or post processing phase, i.e. before carrying out the steps e) to m) at step d) or after having carried out the steps a) to m), the said optimisation of the distribution of the data records in a training and testing set being carried out with the method according to one or more of the preceding claims 1 to 23.

32. The method according to claim 31 comprising the following steps of optimisation:

- defining a set of one or more distributions of the database records onto respective training and testing subsets;

- using the defined set of distributions to train and test a first generation set of one or more prediction algorithms and assigning a fitness score to each;

- feeding the set of prediction algorithms to an evolutionary algorithm which generates a set of one or more second generation prediction algorithms and assigns a fitness score to each; and

- continuing to feed each generational set of prediction algorithms to the evolutionary algorithm until a termination event occurs;

- where said termination event is at least one of a prediction algorithm is generated with a fitness score equalling or exceeding a defined minimum value, the maximum fitness score of successive generational sets of prediction algorithms converging to a given value, and a certain number of generations having been generated.

33. The method according to claims 31 or 32, comprising the following steps:

- generating a population of prediction algorithm each one of them is trained and tested according to a different distribution of the records of the data set in the complete database onto a training data set and a testing data set;

- each different distribution being created by a random or pseudo-random distribution;

- each prediction algorithm of the said population is trained according to its own distribution of records of the training set and is validated in a blind way according its own distribution on the testing set;

- a score reached by each prediction algorithm is calculated in the testing phase representing its fitness;

- an evolutionary algorithm being further provided which combines the different models of distribution of the records of the complete data set in a training and in a testing set which sets are represented each one by a corresponding prediction algorithm trained and tested on the basis of the said training and testing data set according to the fitness score calculated in the previous step for the corresponding prediction algorithm;

- the fitness score of each prediction algorithm corresponding to one of the different distributions of the complete data set on the training and the testing data sets being the probability of evolution of each prediction algorithm or of each said distribution of the complete data set on the training and testing data sets;

- Repeating the evolution of the prediction algorithm generation for a finite number of generations or till the output of the genetic algorithm converges to a best solution and/or till the fitness value of at least some prediction algorithm related to an associated data records distribution has reached a desired value;

- Setting the data records distribution for the best solution as the optimized training and testing subsets for training and testing prediction algorithm.

34. A microarray for genotyping comprising a reduced number of genes, alleles or polymorphisms characterized in that the reduced number of the said genes, alleles or polymorphisms has been selected by means of a method according to claims 30 to 33.